

Preliminary fatty liver disease grading using general-purpose online large language models: ChatGPT-4 or Bard?

To the Editor:

Non-alcoholic fatty liver disease (NAFLD) is a highly prevalent metabolic disorder,¹ pathologically characterized by excessive fat accumulation in the liver. Timely pathological staging and targeted intervention are essential for preventing disease progression and improving outcomes. Liver biopsy is the gold standard for staging NAFLD, but the interpretation requires specialized pathology expertise. In this context, artificial intelligence (AI) may hold promise in aiding the pathological diagnosis of NAFLD, especially for regions where pathology specialists are limited.² As insightfully discussed in a recent review by Vlad Ratziu *et al.* in the *Journal of Hepatology*, digital pathology (DP) and AI approaches could be used to automatically detect and quantify histological features, and may assist pathologists with grading and staging of liver biopsies.³

However, despite the huge potential of DP/AI approaches in this field, for low- and middle-income countries, the need for additional calculating power and fine-tuning of deployment may present a barrier to adoption, especially in underserved regions.⁴ With the rapid advances in generative AI, the large language models (LLMs) including ChatGPT-4 and Google Bard recently introduced image-reading capabilities.⁵ Complemented with their broad user base in global areas including low- and middle-income countries, these generative AIs – if capable of preliminary NAFLD pathological staging – may have significant implications for cost effectiveness and resource allocation. In this exploratory study, we assessed the capability of multimodal ChatGPT-4 and Google Bard to identify the presence/absence of NASH and fibrosis from histological images, as these conditions were more closely associated with subsequent adverse events compared with simple steatosis.⁶

As shown in Fig. 1, between Oct 15th and 25th, 2023, we curated six images (Image 1-6) representing histologically proven NAFLD from the online database PathologyOutlines.⁷ A pathologist independently assessed each image to identify the presence/absence of NASH and grades of fibrosis. These images were labeled as shown in Fig. 1. To avoid chained prompts and to improve robustness, each image was independently presented 20 times to both ChatGPT-4 and Google Bard in separate conversation interfaces with the prompt: “Can you describe this histological image of NAFLD using HE/Azan-Mallory staining? Which NAFLD stage does this image represent?”. This led to 120 responses per AI. Two attending-level hepatologists (YZ, HL) then evaluated these responses, focusing on correctly identifying NASH (images 1 to 4) and the stages of fibrosis (images 5 and 6), considering the different staining methods used in the images. A response was con-

sidered “accurate” when the AI correctly identified the presence or absence of NASH and the stage of fibrosis. The final grading of each response was determined through consensus in panel discussions by both hepatologists. The accuracy of ChatGPT-4 and Google Bard was compared using Pearson's Chi-squared test with Yates' continuity correction.

ChatGPT-4 achieved 87.5% accuracy and Google Bard achieved 38.3% accuracy (Chi-square = 60.1, $p < 0.0001$). ChatGPT-4 consistently maintained a good accuracy across various histopathology images, including those representing simple steatosis (Image 1), NASH (Image 2, 3 and 4), and advanced fibrosis (Image 5 and 6). In contrast, Google Bard's accuracy was notably lower, particularly when identifying images representing advanced fibrosis (Image 5 and 6), with only 3 out of 40 responses being considered accurate. For example, when analyzing image 5, merely two interpretations correctly indicated the presence of advanced fibrosis (F3). Some responses from Google Bard even classified image 5 as absence of fibrosis (supplementary information).

This pilot study revealed that, compared to Google Bard, ChatGPT-4 exhibited a superior proficiency in providing accurate interpretations in preliminary histological analysis of liver biopsies, specifically pertaining to NASH and fibrosis. Our findings highlighted the potential of ChatGPT-4 in aiding preliminary NAFLD grading. As ChatGPT-4 was already widely accessible through online websites in regions where there may be limited pathology expertise, the implications could be far reaching. Where prior investigations predominantly focused on AI's ability to deal with textual content,⁸ our study represents a pioneering effort to evaluate the performance of generative AI in processing histological microscopic images.⁹ More importantly, while another study focused solely on GPT-4, demonstrating just modest diagnostic performances across other organs and tissues,⁹ we for the first time showed that ChatGPT-4 was superior to Google Bard in identifying NASH and fibrosis from histological images. The observed impressive performance of ChatGPT-4 could be attributable to the extensive user feedback, large parameter set, and advanced reasoning abilities.¹⁰ This study has limitations. First, to ensure reproducibility and consistency, our study used histological images from an open online database. However, the collection of a limited fragment of the liver is not representative of the entire liver and may not be able to capture the variability of fibrosis distributions, which may lead to misinterpretation by pathologists. Second, due to relatively limited sample size, more studies are warranted to validate our findings. Generative AI with LLMs – especially ChatGPT-4 – may offer further opportunities to utilize DP/AI approaches.

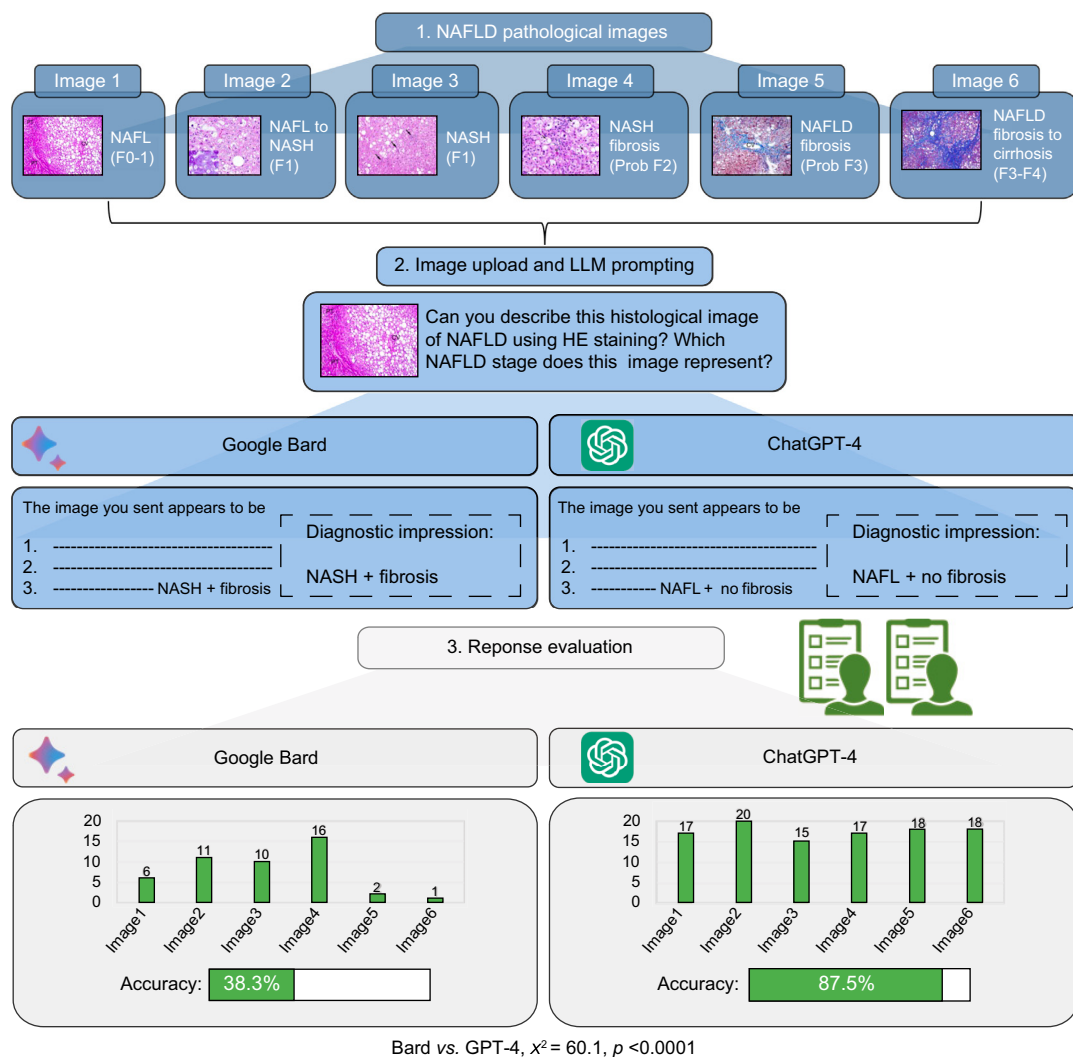


Fig. 1. Workflow and performance of ChatGPT and Google Bard in NAFLD staging using histological images. (1) This diagram illustrated that we curated 6 histological images from an online database for clinical pathology, with extensive links to images including NAFLD and NASH. These images were contributed by Yoshihiro Ikura, M.D. (<https://www.pathologyoutlines.com/topic/liverNASH.html>) and a pathologist independently evaluated each image to identify the presence/absence of NASH and grades of fibrosis: Image 1 (HE staining, nonalcoholic fatty liver or simple steatosis, F0-1), Image 2 (HE staining, simple steatosis to NASH, F1), Image 3 (HE staining, NASH, F1), Image 4 (HE staining, NASH and fibrosis, probable F2), Image 5 (Azan-Mallory staining, fibrosis, probable F3), and Image 6 (Azan-Mallory staining, fibrosis to cirrhosis, F3-4). (2) The images were posed to both ChatGPT and Bard with the prompt “Can you describe this histological image of NAFLD using HE staining/Azan-Mallory staining? Which NAFLD stage does this image represent?”. (3) Responses and diagnostic impressions from ChatGPT and Bard were recorded and then evaluated by two independent hepatologists. We focused on their capabilities to correctly identify NASH (image 1 to 4) and fibrosis stages (image 5 and 6), considering the different staining methods used in the images. The accuracy of ChatGPT-4 and Google Bard was compared using Pearson’s Chi-squared test with Yates’ continuity correction. NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis.

Yiwen Zhang^{1,†}
Hanyun Liu^{2,†}
Bin Sheng^{3,#}
Yih Chung Tham^{4,#}
Hongwei Ji^{5,6,*,#}

¹Department of Endocrinology and Metabolic Hepatology, The Affiliated Hospital of Qingdao University, Qingdao, China

²Department of Infectious Disease and Hepatology, The Affiliated Hospital of Qingdao University, Qingdao, China

³Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

⁴Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

⁵Tsinghua Medicine, Tsinghua University, Beijing, China

⁶Department of Internal Medicine, Beijing Tsinghua Changgung Hospital, Beijing, China

*Corresponding author. Address: Tsinghua Medicine, Tsinghua University, Beijing, China.

E-mail address: hongweijcn@gmail.com (H. Ji).

Received 13 November 2023; Accepted 17 November 2023; Available online 28 November 2023
<https://doi.org/10.1016/j.jhep.2023.11.017>

© 2023 European Association for the Study of the Liver. Published by Elsevier B.V. All rights reserved.

[†]Contributed Equally

[#]Senior authors



Financial support

This study was funded in part by the National Key R & D Program of China (2022YFC2502800), National Natural Science Foundation of China (82103908), the Shandong Provincial Natural Science Foundation (ZR2021QH014), the Shuimu Scholar Program of Tsinghua University, and National Postdoctoral Innovative Talent Support Program (BX20230189). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflicts of interest

The authors of this study declare that they do not have any conflict of interest. Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Acquisition of data: Yiwen Zhang, Hanyun Liu, Hongwei Ji; Analysis and interpretation of data: All authors; Drafting of the manuscript: All authors; Critical revision of the manuscript for important intellectual content: All authors; Statistical analysis: Yiwen Zhang, Hongwei Ji; Obtained funding: Hongwei Ji; Study supervision: Hongwei Ji.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhep.2023.11.017>.

References

Author names in bold designate shared co-first authorship

- [1] Devarbhavi H, Asrani SK, Arab JP, et al. Global burden of liver disease: 2023 update. *J Hepatol* 2023;79:516–537.
- [2] Sayed S, Cherniak W, Lawler M, et al. Improving pathology and laboratory medicine in low-income and middle-income countries: roadmap to solutions. *Lancet* 2018;391:1939–1952.
- [3] Ratziu V, Hompesch M, Petitjean M, et al. Digital pathology and artificial intelligence in non-alcoholic steatohepatitis: current status and future directions. *J Hepatol* 2024;80(2):335–351.
- [4] Mollura DJ, Culp MP, Pollack E, et al. Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology* 2020;297:513–520.
- [5] ChatGPT. ChatGPT. 2023. Release Notes, <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>.
- [6] Sanyal AJ, Harrison SA, Ratziu V, et al. The natural history of advanced fibrosis due to nonalcoholic steatohepatitis: data from the simtuzumab trials. *Hepatology* 2019;70:1913–1927.
- [7] Ikura Y. Nonalcoholic fatty Liver disease/nonalcoholic steatohepatitis (NASH). <https://www.pathologyoutlines.com/topic/liverNASH.html>. PathologyOutlines.com website Accessed October 15th, 2023.
- [8] Zhang Y, Wu L, Mu Z, et al. Assessing state-of-the-art online large language models for patient education regarding steatotic liver disease. *Clin Mol Hepatol* 2024;30(1):113–117.
- [9] Chaoyi Wu JL, Zheng Qiaoyu, Zhao Weike, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. *arXiv:2310.09909v2* 2023.
- [10] OpenAI. GPT-4 technical report. *arXiv:2303.08774* 2023.